(2)

RSRE
MEMORANDUM No. 4359

ROYAL SIGNALS & RADAR ESTABLISHMENT

EXPERIMENTS WITH GRAND VARIANCE IN THE ARM CONTINUOUS SPEECH RECOGNITION SYSTEM

Authors: M J Russell & K M Ponting

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
R S R E MALVERN,
WORCS.

R S R E MEMORANDUM No. 4359

UNLIMITED

CONDITIONS OF RELEASE

0066657 BR-113304

********** DRIC U

********** DRIC Y

Reports quoted are not necessarily available to members of the public or to commercial organisations.

Royal Signals and Radar Establishment
Memorandum 4359

Experiments with Grand Variance in the ARM Continuous Speech Recognition System

M J Russell and K M Ponting
Speech Research Unit, SP4,
Royal Signals and Radar Establishment,
St. Andrews, Great Malvern, England

8th February 1990

Abstract

The use of triphones to cope with contextual effects in phoneme-level hidden Markov model (HMM) based speech recognition results in a huge increase in the number of system parameters which need to be estimated. The solution to this problem is to reduce the number of independent system parameters so that those which remain can be estimated more robustly from the training data. For HMMs with Gaussian state output probability density functions (pdfs), a simple example of such an approach is the "grand" variance method in which all state output pdfs share the same covariance matrix. This paper reports the results of experiments designed to investigate the effect of grand variance on the performance of the triphone-HMM based *ARM* continuous speech recognition system.

1 Introduction

The work described in this research note was conducted at the UK Speech Research Unit as part of the Airborne Reconnaissance Mission (*ARM*) continuous speech recognition project. The aim of the *ARM* project is accurate recognition of continuously spoken airborne reconnaissance reports using a speech recognition system based on phoneme-level hidden Markov models (HMMs). The *ARM* project is described in [2]. The work described here is based on version 5 of the *ARM* system.

The more recent versions of the *ARM* system use triphone HMMs to model the context-sensitivity of the acoustic patterns corresponding to phonemes. This approach makes the simplifying assumption that context-related variations in the acoustic realisation of a particular phoneme depend only on the immediately preceding and following phonemes. This means that rather than modelling a phoneme using a single HMM, each phoneme is modelled using a set of HMMs, one for each pair of phonemes which occur as its immediate neighbours in the *ARM* baseform dictionary.

Depending on the speaker, there are approximately 1500 word-internal triphones in the *ARM* vocabulary, resulting in a speech recognition system with approximately 234,000 parameters. Assuming that 20 minutes of speech is used to train the system, the number of training observations is 3,120,000, or approximately 13 observations per parameter. These observations are not statistically independent, nor are they uniformly distributed between triphones. In fact approximately 400 of the triphones in the *ARM* vocabulary are not represented in the training set. Consequently many of the triphone HMM parameters will be undertrained.

The solution to this training problem is to reduce the number of independent system parameters so that those which remain can be estimated more robustly from the training data. The most obvious way to achieve this is to "tie" together different system parameters so that they share the same training material. The simplest example of such an approach is the "grand" variance method [3] in which all HMM state output probability density functions share the same covariance matrix. This note reports the results of applying the grand variance method in the context of the *ARM* system.

2 The Triphone Based *ARM* system (*ARM-5*)

The version of the *ARM* system which is used in the present experiments is *ARM-5* (see [2] for a description of the evolution of the *ARM* system).

Front-end acoustic analysis in all versions of the *ARM* system is derived from the SRUbank filterbank analyser in its default configuration of 27 critical band filters

spanning the range 0 to 10kHz and producing 100 frames per second. In the present experiments two alternative front-end representations were used. These are refered to as *CC16* and *CC12 δ* ([4]), and are derived as follows.

Let $\vec{v_t} = (v_t{}^1, ..., v_t{}^{27})$ be the SRUbank feature vector at time t. The mean channel amplitude $m(\vec{v_t})$ of $\vec{v_t}$ is subtracted from each component of $\vec{v_t}$, and the resulting vector is then rotated using a discrete cosine transform to obtain a new feature vector $\vec{w_t}$. The 17 dimensional feature vector $\vec{x_t}$ for representation *CC16* at time t is defined by:

$$x_t{}^d = w_t{}^d,\ d = 1, ..., 16$$
$$x_t{}^{17} = m(\vec{v_t})$$

and the 26 dimensional feature vector $\vec{y_t}$ for parameterisation *CC12 δ* is given by:

$$y_t{}^d = w_t{}^d,\ d = 1, ..., 12$$
$$y_t{}^{13} = m(\vec{v_t})$$
$$y_t{}^d = (w_{t+2}{}^d - w_{t-2}{}^d),\ d = 14, ..., 25$$
$$y_t{}^{26} = (m(\vec{v_{t+2}}) - m(\vec{v_{t-2}}))$$

Detailed results of experiments which have been conducted to assess the performance of a range of related front-end representations derived from linear transformations of SRUbank are presented in [4].

Acoustic-phonetic processing in *ARM-5* uses a set of approximately 1500 HMMs (the precise number depends on the speaker) consisting of:

- Four single state "non-speech" HMMs to cope with non- speech sounds in regions of the test data between spoken sentences.
- Six word-level HMMs for the commonly occuring short words "air", "at", "in", "of", "oh" and "or". The number of states in each of these word-level HMMs is equal to three times the number of phonemes in the baseform transcription of the corresponding word.
- Approximately 1490 three-state triphone HMMs, one for each word- internal triphone which occurs in the *ARM* vocabulary. Since the baseform pronunciations of *ARM* vocabulary words vary between speakers in the speaker dependent *ARM* system, the precise number of triphone HMMs will be different for each speaker.

As with earlier versions of the *ARM* system, all HMM states in *ARM-5* are identified with single multivariate Gaussian state output probability density functions with diagonal (co)variance matrices.

Words in the *ARM* vocabulary are related to phonemes through a dictionary of "baseform" phonemic transcriptons. In the current, speaker-dependent, version

of the *ARM* system this dictionary is modified for each speaker. These modifications are concerned with broad differences, for example between "northern english" and "southern english", rather than with fine details of the speakers pronunciation. It is assumed that spoken examples of vocabulary words conform to these baseform transcriptions.

3 HMM Training and Recognition

3.1 Training and Test Data

Speaker dependent recognition experiments were conducted using speech from a single speaker (SJ) as training and test material. The training set consisted of 37 *ARM* reports (224 sentences, 1985 words) chosen to give maximum coverage of phonemes which occur infrequently in the *ARM* vocabulary. Ten reports from the same speaker (540 words, 2293 phonemes according to baseform transcriptions) were used as test material.

3.2 Monophone HMM Training

Initial estimates of the parameters of context-insensitive monophone phoneme HMMs were obtained from the equivalent of two *ARM* reports of speech, hand labelled at the phoneme level. Similarly, initial estimates of the common word HMM parameters were obtained from single examples of these words extracted from continuous speech. The initial estimates of parameters of a single state "non-speech" HMM were derived from a typical non-speech region of the training data. This model was used as the initial model for all four non-speech HMMs. The models were optimised with respect to the complete training set labelled orthographically at the sentence level. Standard sub-word HMM training procedures were used in which sentence level HMMs were constructed from phoneme-level HMMs using the dictionary of baseform transcriptions of *ARM* vocabulary words. These models were then mapped onto the sentence level acoustic data using the forward backward algorithm to obtain contributions to the model parameter estimates.

3.3 Triphone HMM Training

The parameters of the context insensitive monophone HMMs were used as the initial estimates for the parameters of the set of triphone HMMs. The triphone HMMs were then optimised with respect to the complete training set labelled orthographically at the sentence level using the standard sub-word HMM training procedures.

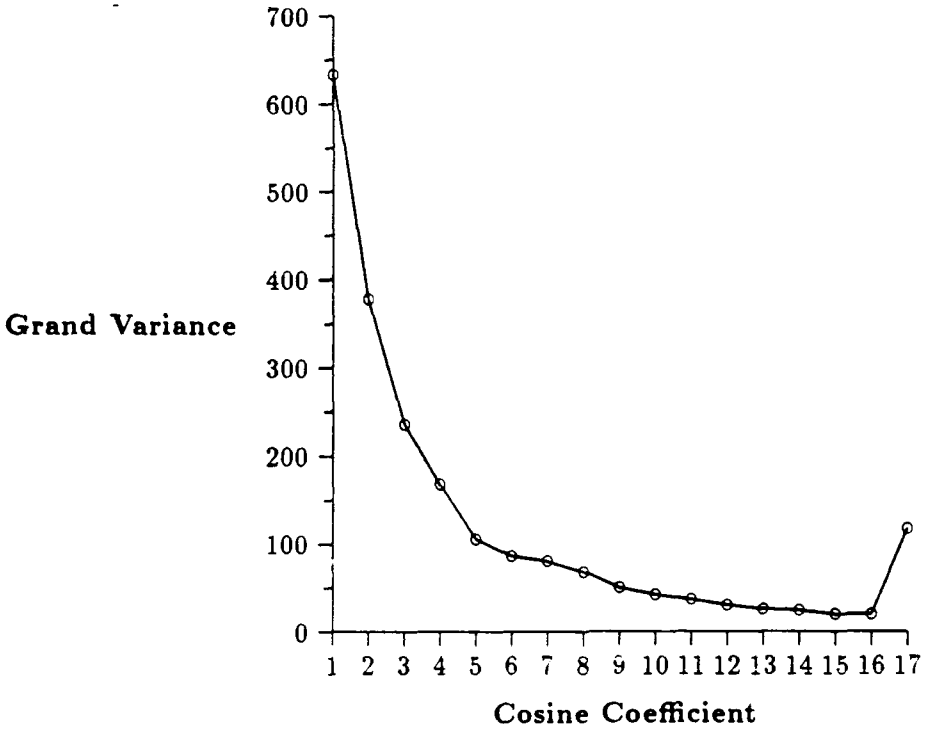

Figure 1: *Grand variance as a function of component of the* CC16 *front-end representation.*

3.4 Estimation of Grand Variance

The grand diagonal (co)variance matrix was estimated using a further pass of the training algorithm applied, as above, to the complete training set labelled orthographically at the sentence level. During this stage of training all other parameters were fixed. This training scheme will be refered to as *GV-1*.

It was found to be beneficial to use two further iterations of the training algorithm: the first to reestimate the mean vectors of the state output pdfs given the grand diagonal covariance matrix, and the second to do a final reestimation of the grand covariance matrix. This scheme will be refered to as *GV-2*.

Figure 1 shows grand variance as a function of the components of the *CC16* parameterisation. As one would expect ([4]) most of the variance is concentrated in the lower-order components. Notice that the variance increases for the 17^{th} component because in the *CC16* parameterisation this component is the mean SRUbank channel amplitude and not a cosine coefficient.

3.5 Recognition

Recognition was performed using a one-pass dynamic programming algorithm with beam search and partial traceback [1]. Results are presented in terms of *% words (or phonemes) correct* and *% word (or phoneme) accuracy*. These are computed as follows, using dynamic programming to align the true transcription of the test data with the output of the recogniser:

$$\%\ words\ correct = \frac{N - S - D}{N} \times 100,$$
$$\%\ word\ accuracy = \frac{N - S - D - I}{N} \times 100$$

where N is the number of words in the test set, and S, D and I are the number of words recognised as the incorrect word, deleted and inserted respectively.

Four different syntaxes were used to constrain the recognition process: a *word* syntax, which allows recognition of any sequence of words from the ARM vocabulary; a *full* syntax (perplexity 6) which was used to generate the ARM reports, a phoneme based *simple* syntax which allows any sequence of phonemes to be recognised, and a phoneme based *trisimple* syntax which forces the recogniser to consider only sequences of triphone HMMs which are consistent in the sense that the triphone $(a : b\_c)$, corresponding to the phoneme a preceeded by b and followed by c, can only be preceded and followed by triphones of the form $(b : *\_a)$ and $(c : a\_*)$ respectively, where $*$ denotes an arbitrary phoneme or word boundary symbol.

4 Experiments and Results

Tables 1 and 2 show the results of phoneme and word recognition experiments respectively for the *CC16* front-end representation. Tables 3 and 4 show the corresponding results for the *CC12 δ* front-end. Results for context-insensitive monophone HMMs are included for comparison.

The results show that the effect of grand variance on phoneme recognition is quite different to its effect on word recognition. They also suggest that the dimensionality of the acoustic front-end parameterisation is an important factor. Word recognition and phoneme recognition will be considered separately.

4.1 Word Recognition Results

The discussion of the word recognition results will concentrate on % word accuracy with no syntax.

	Phoneme Syntax (perplexity=47)		Trisimple Syntax	
Training Scheme	Phonemes Correct	Phoneme Accuracy	Phonemes Correct	Phoneme Accuracy
Monophones	64.3%	47.1%	-	-
Triphones	84.3%	58.7%	90.0%	85.2%
GV-2	84.5%	51.9%	-	-

Table 1: Results of phoneme recognition experiments using the *CC16* parameterisation (540 word test set).

	Word Syntax (perplexity=497)		Full Syntax (perplexity=6)	
Training Scheme	Words Correct	Word Accuracy	Words Correct	Word Accuracy
Monophones	81.5%	55.7%	98.3%	97.0%
Triphones	86.5%	66.5%	92.4%	86.9%
GV-2	96.3%	86.1%	99.4%	99.3%

Table 2: Results of word recognition experiments using the *CC16* parameterisation (540 word test set).

The results suggest that the effect of moving from a monophone to a triphone based sytem with state specific covariance matrices depends on the dimensionality of the acoustic front-end. In the case of the 17 dimensional *CC16* representation, word accuracy with no syntax rises from 55.7% to 66.5%. By contrast, with the 26 dimensional *CC12* δ representation, performance falls from 52.2% for monophones to 37.0% for triphones. This result suggests that the training set cannot support the increased number of parameters in the *CC12* δ based *ARM* system.

For both front-end representations, the introduction of grand variance leads to substantial improvements in word recognition accuracy relative to both monophone HMMs and triphone HMMs with state-specific covariance matrices. The performances of the monophone, triphone and *GV-2* systems are 55.7% , 66.5% and 86.1% for the *CC16* front-end, and 52.2%, 37.0% and 81.3% for the *CC12* δ front-end.

It can also be seen from the rows of table 4 labelled *GV-1* and *GV-2* that the adjustment of the state means relative to the first estimate of grand variance, and the subsequent reestimation of grand variance (see section 3.4) leads to a useful increase in recognition accuracy.

	Phoneme Syntax (perplexity=47)		Trisimple Syntax	
Training Scheme	Phonemes Correct	Phoneme Accuracy	Phonemes Correct	Phoneme Accuracy
Monophones	66.2%	53.3%	-	-
Triphones	88.9%	71.5%	92.1%	86.6%
GV-1	89.4%	53.5%	96.2%	89.7%
GV-2	90.1%	60.4%	96.7%	91.8%

Table 3: Results of phoneme recognition experiments using the *CC12 δ* parameterisation (540 word test set).

	Word Syntax (perplexity=497)		Full Syntax (perplexity=6)	
Training Scheme	Words Correct	Word Accuracy	Words Correct	Word Accuracy
Monophones	79.8%	52.2%	99.1%	98.7%
Triphones	73.5%	37.0%	89.3%	83.0%
GV-1	94.4%	78.5%	99.4%	99.1%
GV-2	94.8%	81.3%	99.4%	99.1%

Table 4: Results of word recognition experiments using the *CC12 δ* parameterisation (540 word test set).

4.2 Phoneme Recognition Results

The results of the experiments in phoneme recognition are quite different from those for word recognition. Phoneme recognition accuracy is significantly better for triphone HMMs with state-specific covariance matrices than for context-insensitive monophone HMMs. For example, in the case of the *CC12 δ* parameterisation phoneme recognition accuracy with the phoneme syntax is 53.3% for monophones and 71.5% for triphones with state-specific covariance matrices. Furthermore, and in contrast with the results for word recognition, the use of grand variance consistently results in a significant drop in phoneme recognition accuracy (without syntax) relative to triphone HMMs with state-specific covariance matrices. Using the *CC12δ* parameterisation again as an example, phoneme accuracy drops from 71.5% to 60.4% when state specific covariance matrices are replaced with a grand covariance matrix.

4.3 Discussion

The superior performance at the phoneme level of triphone HMMs without grand variance over monophone HMMs suggests that the use of several (possibly undertrained) HMMs to model the acoustic realisation of a phoneme is better from the viewpoint of discrimination than a single HMM. The fact that these models can lead to a fall in word recognition accuracy (as is the case with the *CC12 δ* parameterisation) suggests that when a phoneme recognition error does occur it is too severe to be corrected by the word syntax. The hypothesis that the system is making relatively "hard" decisions at the phoneme level is consistent with the use of possibly undertrained state-specific covariance matrices.

The use of a grand covariance matrix has the effect of "softening" decisions at the phoneme level. This softening is clearly too extreme for accurate phoneme recognition and results in poorer phoneme recognition accuracy. However it increases the relative importance of the word syntax and in this way leads to improved word recognition accuracy.

5 Conclusions

The experiments described in this research note demonstrate that the use of a grand covariance matrix is critical for the high word recognition accuracies which have been demonstrated by the triphone based *ARM* system. However, the gain in performance relative to context insensitive monophone HMMs is not a consequence of improved recognition accuracy at the phoneme level, since phoneme accuracy is actually made worse by the use of grand variance. Rather, it is a consequence of an improved balance between the scores which are derived from the acoustic models and the constraints of the word syntax.

References

[1] J S Bridle, M D Brown and R M Chamberlain, "A one-pass algorithm for connected word recognition", IEEE-ICASSP, 899-902, 1982.

[2] M J Russell, K M Ponting, S M Peeling, S R Browning, J S Bridle and R K Moore, "The ARM Continuous Speech Recognition System", Proc. ICASSP'90, Albuquerque, New Mexico, April 1990.

[3] D B Paul, "A speaker-stress resistant isolated word recognizer", ICASSP'87, Dallas, TX, 1987.

[4] M J Russell, D Lowe, M D Bedworth and K M Ponting, "Improved Front-End Analysis in the ARM System: Linear Transformations of SRUbank", RSRE Memorandum Number 4358, 1990.

REPORT DOCUMENTATION PAGE

DRIC Reference Number (if known) .....

Overall security classification of sheet .....Unclassified.....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the field concerned must be marked to indicate the classification eg (R), (C) or (S).

Originators Reference/Report No.	Month	Year
MEMO 4359	FEBRUARY	1990

Originators Name and Location
RSRE, St Andrews Road
Malvern, Worcs WR14 3PS

Monitoring Agency Name and Location

Title
EXPERIMENTS WITH GRAND VARIANCE IN THE ARM CONTINUOUS SPEECH RECOGNITION SYSTEM

Report Security Classification	Title Classification (U, R, C or S)
Unclassified	U

Foreign Language Title (in the case of translations)

Conference Details

Agency Reference	Contract Number and Period

Project Number	Other References

Authors	Pagination and Ref
Russell, M J; Ponting, K M	9

Abstract

The use of triphones to cope with contextual effects in phoneme-level hidden Markov model (HMM) based speech recognition results in a huge increase in the number of system parameters which need to be estimated. The solution to this problem is to reduce the number of independent system parameters so that those which remain can be estimated more robustly from the training data. For HMMs with Gaussian state output probability density functions (pdfs), a simple example of such an approach is the "grand" variance method in which all state output pdfs share the same covariance matrix. This paper reports the results of experiments designed to investigate the effect of grand variance on the performance of the triphone-HMM based *ARM* continuous speech recognition system.

Abstract Classification (U,R,C or S)
U

Descriptors

Distribution Statement (Enter any limitations on the distribution of the document)

Unlimited